To do list:

1. Methods er bade I nutid og datid. Ret det til det rigtige??
2. Provide graphs and plots and flow charts for methods section
3. Read through and see if any ‘notes to section’ should be included in the sections, depending on space. This in OneNote!
4. Write about “software used” somewhere in the methods section
5. Explain why random forest/cvsm/others is good in the paper.
6. Remember to evaluate accuracy not to chance but to majority group
7. Make sure Pipeline contains information on moving away from “accuracy”
8. Check if I need more in introduction on the specifics of my machine learning
9. Mention software VSCODE, Python, etc.
10. Mention all relevant places that: Even real diagnoses might be different (in discussion) – if thought of as continuum, perhaps only more schizophrenic people are diagnosed in Asia?? Or perhaps people diagnose differently across countries?
11. Read through papers in “papers downloaded” and add throughout where needed.
12. Skriv ML efter første gang det bliver nævnt, i stedet for machine learning
13. Should machine learning method (SVM linear kernel) be mentioned in detail?
14. What an ensemble model is and why we’re using it should be incorporated in the text
15. Nævn i methods et sted at vi arbejdede med predictions på voice-fil niveau og ikke participant!
16. Uddyb i methods at vi laver ”stacking” ensemble predictions. Så det bliver udspecificeret
17. Acknowledgements add or no? Perhaps also depends on Table of Contents (does it fit or not)
18. Can it really by utilized in the real world? (This model will believe that roughly 50% of population is schizo) – not going into depth with this.
19. Det originale studie ser ud som om det tester within-sample, mens mit tester out-of-sample. Skal jeg lade være med at nævne det? Eller skal jeg skrive omkring det?

# Abstract

Can machine learning (ML) applied to voice data be used as a tool to help diagnose people with schizophrenia? Numerous studies have shown high ML accuracy when classifying schizophrenia, but the ways in which they do so differ widely, as concluded in the latest meta study within the field. Little work has been done to replicate these previous ML methods on new data, and there is currently no consensus on which methods should be used.  
This study replicated two promising ML studies on new data, using an improved validation technique and an inclusion of sensitivity and specificity rates. Accuracy rates found through replication were dissimilar to the original studies, with study X\* and study Y\* having overall accuracy rates for classification at 60% and 67%. In other words, a drop of 6 and 3 percentage points for the two studies, respectively. Through discussion, this study has found that the difference in scores in the replication points toward low ecological validity and robustness. The rest of the literature was also discussed, and I found that the widely heterogeneous results within the field indicate similar trends.  
As a consequence, this study has attempted to establish a ML pipeline less prone to the pitfalls of ML, with the intention of establishing a general procedure for future research. Finally, this paper advocates for a more open and cumulative scientific community.

How to write abstract:  
<https://blackboard.au.dk/bbcswebdav/pid-2793891-dt-content-rid-9152972_1/courses/BB-Cou-Hold-36086/L1%20-%20Getting%20started.pdf>  
p. 14 - p. 18

# 1. Introduction

## 1.1 Schizophrenia and biomarkers

### 1.1.1 Schizophrenia

**Actual paper:**

**Notes for section:**

Biomarkers - why voice good? -> shows part of social impairment.

### 1.1.2 Biomarkers and voice atypicalities

**Actual paper:**  
  
**Notes for section:**

Biomarkers - why voice good? -> shows part of social impairment.

1. Schizophrenia and voice in general
   1. 3 methods of studying
2. Two meta-studies -> the lit. is a mess

1.

History of the project:

Schizophrenia has certain distinctive features vocally. (Alogia, blunt affect, "poverty of speech", "latency of speech" etc.). This has been known since forever (Bleuler, 1911; Kraepelin, 1919).

 Voice atypicalities in SZ’s have always been known (Bleuler, 1911; Kraepelin, 1919).  
Schizophrenia has certain distinctive features vocally. Qualitatively the atypicalities have been described using numerous different terms (Alogia, blunt affect, "poverty of speech", "latency of speech", increased pauses, distinctive tone, intensity of voice etc.).

Voice atypicalities have been studied using 3 methods. Qualitative perceptual ratings, quantitative acoustic analysis and ML investigations.

Qualitative perceptual ratings have found robust differences between SZ and TD.

Quantitative acoustic analyses have found fewer robust differences, with varying effect sizes and sometimes direction.

Quantitative acoustic analyses have identified acoustic features on the basis of automated processes, leaving the assessment of the acoustic features more reliable. Using automation, the features of a set of voice data will identical over multiple feature detections, given the same feature detection hard- and software.  
Here, fewer robust differences were found with varying effect sizes and direction, depending on the features investigated (Cohen et al., 2014; https://www.biorxiv.org/content/10.1101/583815v4.full.pdf).

This is also in next section?:

Multivariate ML investigations have found promising results. Focus on minimizing out-of-sample-error instead of within sample-error as when using more traditional analyses, makes the applicability of the method more practically generalizable. It also allows for analyzing multiple features in conjunction. High correlation between almost all features (3.3, correlation <https://www.biorxiv.org/content/10.1101/583815v4.full.pdf> ).  
It does, however, not allow for transparency as to wherein the acoustic differences between SZ and HC lie.

2.

We don't know which features proves to have differences between SZ and TD

The litt. is a mess - results in different directions.

There's already a metastudy on Schizophrenia; which found atypicalities on different voice/speaking parameters - with varying effect sizes.

Large heterogeneity between studies.

More demanding tasks meant larger effect sizes.

## 1.2 Machine learning for detection of acoustic patterns

### 1.2.1 Prospects of machine learning in classifying schizophrenia

**Actual paper:**  
  
**Notes for section:**

1. Metatext and motivation for going into depth with machine learning
2. Allows for finding features (feature selection)
3. Allows for analyzing multiple features in conjunction
4. Promising findings (high accuracy in many studies)
5. Less interpretability but more practical applications (cheap)

Will not go into other ML things (gesticulation or others – beyond the scope of this paper)

1.

Although:  
Qualitative perceptual ratings have found relatively robust differences in voice between SZ and TD. Relying on raters to assess perceptual differences has some limitations. A feature such as “latency of speech” is interpretable and is partly going to be rated on the basis of human intuition – this requires comprehensive training for the rater. Moreover, the complex interplay between multiple acoustic features is hardly very accessible, even given proper and rigorous training.  
  
Therefore:

ML doesn’t have this problem.

2.

Feature selection; ridge, lasso, elasticnet

3.

Multivariate ML investigations have found promising results. Focus on minimizing out-of-sample-error instead of within sample-error as when using more traditional analyses, makes the applicability of the method more practically generalizable. It also allows for analyzing multiple features in conjunction. High correlation between almost all features (3.3, correlation <https://www.biorxiv.org/content/10.1101/583815v4.full.pdf> ).

4.

Multivariate ML investigations have found promising results. Focus on minimizing out-of-sample-error instead of within sample-error as when using more traditional analyses, makes the applicability of the method more practically generalizable. It also allows for analyzing multiple features in conjunction. High correlation between almost all features (3.3, correlation <https://www.biorxiv.org/content/10.1101/583815v4.full.pdf> ).

5.

Clinical application -> given schizophrenia, and given samtaleterapi or drugs, see how they're doing along the way by them talking every week on their phone.  
"*In addition, voice analysis may potentially allow to assess the response to psychosocial or pharmacological treatment over longer periods using objective and quantitative indices, and enhance the capability of clinicians to capture the complex relationship between emotion regulation, expressive behavior, social perception and cognitive and clinical features of the disorder (e.g. Ben-Zeev et al., 2017; Dahlgren et al., 2018; Tahir et al., 2019)*" (Parola, Fusaroli et. al 2019)Va [(Bush et al., 1998)](https://www.zotero.org/google-docs/?KFKj12).  
  
It does, however, not allow for transparency as to wherein the acoustic differences between SZ and HC lie.

ML can perhaps help with showing:  
a) Severity of schizophrenic symptoms  
b) Diagnosis, schizophrenia

Practical applications:   
Assisting tool for assessing diagnosis (Parola, Fusaroli et. al 2019)

On prediction of severity of clinical features from acoustic measures:  
(Püschel et al., 1998)

6 (extra):

Applicability of Bachelors project:

Meta-science, open science.

Assisting tool for assessing diagnosis (Parola, Fusaroli et. al 2019)

Clinical application -> given schizophrenia, and given samtaleterapi or drugs, see how they're doing along the way by them talking every week on their phone.

"*In addition, voice analysis may potentially allow to assess the response to psychosocial or pharmacological treatment over longer periods using objective and quantitative indices, and enhance the capability of clinicians to capture the complex relationship between emotion regulation, expressive behavior, social perception and cognitive and clinical features of the disorder (e.g. Ben-Zeev et al., 2017; Dahlgren et al., 2018; Tahir et al., 2019)*" (Parola, Fusaroli et. al 2019)

Companies interested in this (Lasse Hansen), Switzerland Internship on this in depression

### 1.2.2 Current limitations in the literature

**Actual:**

**Notes:**

* Overfitting
* Differences in methods, method quality and levels of transparency
  + Effect sizes of acoustic features is partially determined by task (difficulty)
* Lack of replications
  + Promising results
  + No validation between datasets
  + Language differences
  + No performance robustness measures

## 1.3 Alleviating current limitations

### 1.3.1 Through replications

**Actual paper:**  
  
**Notes for section:**

s

### 1.3.2 Through proper ML implementation

**Actual paper:**  
  
**Notes for section:**

1. Meta text to have a rød tråd
2. Establish the need for a change in research (from above section)
3. Idea of a pipeline
4. Explain how a pipeline might alleviate the problem. And cause the change that is needed

Set up a standard pipeline.  
(Very) general introduction to pipeline in introduction. Just the conceptual structure.  
  
In intro:

“explain the most basic and most important steps”

BUT … (not feature selection on full dataset) etc.

Mention confusion matrix

In methods:  
Take the individual conceptual steps from the intro, and explain in specific detail how we tailored this ML “to the conceptual steps”.

The proposed pipeline follows a relatively simple overall structure, which can be thought of as 8 steps. It is important to note, that the course of action within each step to a certain extent depends upon the specific data and classification problem. Moreover, it is important to note that documenting extensively throughout is important for increasing transparency. High transparency a) allows for future replications and b) increases the applicability of the work, since knowledge will be available about exactly under which conditions a given machine learning approach performs a certain way. The steps of the pipeline are as follows:   
**1) Data acquisition.** Acquiring the voice data for fitting a machine learning algorithm – this may be data that has been acquired either through recording, or from using already recorded data. **2) Data preprocessing.** Consists of cleaning the data; cutting away irrelevant speech (e.g. from interlocutors), noise removal and extracting features. It might also include applying data augmentation (e.g. applying noise, reverb etc.). **3) Data partitioning.** Partitioning the data into a training and a holdout set for testing. Typically, a split of roughly 80/20 is used. Generally, a larger holdout set means more precise knowledge of the performance of a given model. A larger training set on the other hand generally means better and more robust predictions. **4) Feature scaling and selection.** Feature scaling is necessary for all machine learning algorithms that are distance-based. In these cases, scale the train and holdout set separately. Use information from the training set (e.g. min. and max. values in min./max. normalization) to do the scaling on both sets. This ensures no leakage of information from the training set to the test set, while still providing a common scale without losing information or distorting the differences in the range of values. Feature selection entails selecting only the relevant machine learning parameters, that improves classification. Many different methods can be applied in order to obtain a relevant feature set. It may also be skipped if theory or some motivation dictates it. However, feature selection tends to generally improve the robustness of predictions, shorten training times, avoid the curse of dimensionality and it increases interpretability. **5), 6), 7) Model training, parameter tuning, model testing.** These three steps are intertwined, and the cycle may be repeated. Divide the training set up into training and test and then train and test the model. After seeing how it performs on the test set, you may tune the parameters and repeat the process until the predictions are optimal. This can be done advantageously by the implementation of cross validation. **8) Validation on hold-out set.** The model then predicts the holdout set and performance is evaluated using relevant metrics. To allow for further insights into the performance of a given model, evaluation metrics ought also to be calculated for subgroups of participants (e.g. men/women or nationalities).

An overview of the pipeline can be seen in figure 1.

Fig 1.

*Pipeline with workflow that is well suited for classification machine learning on voice.*

**Notes for section:**

* S

s

### 1.3.3 Thesis statement / purpose of paper

**Actual paper:**  
  
**Notes for section:**

Noone should just mindlessly replicate, when replicating!! Important to take all the necessary steps.

Overall goal of thesis:  
How does it work to reproduce?  
Increase conservative  
What are the important lessons we’re learning while we’re replicating

**Riccardos words on overall goal of thesis:**

How does it work to replicate the results of these paper when you increase the conservativeness of the procedure - and what do you learn from the problems, that future researcher can use when they do these kind of analysis

**Thesis statement idea 1 (Maries):**

This thesis aims to investigate the capabilities of existing machine-learning classifying individuals with ASD from acoustic features. We will review previous literature, extract strong voice-features and machine-learning models, and validate models on new data. We predict that support vector machine will achieve higher accuracy but will have less x and that naive bayes will x. Additionally, we predict that validation methods x,y,z will make results stronger in specific case/weaker generalization. By this, we will attempt to establish a procedure for machine-learning studies that achieve the most robust and ecologically valid measures.

**Thesis statement idea 2:**

This thesis aims to replicate two promising findings of machine learning classification of schizophrenia, using voice data. Since the literature on the area has very heterogeneous findings, I expect worse performance given the new data that I will test on. Given the inrobustness and low ecological validity of ML attempts, I will attempt to establish a ML pipeline less prone to the pitfalls of ML, with the intention of establishing a general procedure for future research.

Tour de Bachelors:

[https://docs.google.com/document/d/1qc3tDtAg6sc2-zfnxaxqKl\_WydK3AAgaDLb1V7QjTTU/edit?fbclid=IwAR1JB53UmJcDEI8GnXEEvA4PcuWXvVeX\_ZN43VEamHHxMWsHYdAR\_Wo3vKY#](https://docs.google.com/document/d/1qc3tDtAg6sc2-zfnxaxqKl_WydK3AAgaDLb1V7QjTTU/edit?fbclid=IwAR1JB53UmJcDEI8GnXEEvA4PcuWXvVeX_ZN43VEamHHxMWsHYdAR_Wo3vKY)

Pipeline:

[https://docs.google.com/document/d/1fbfpR5ZQiVTZYChzWut06fkXA9CQziMMJNtFOiF6Giw/edit?fbclid=IwAR3MU3OTehQ\_nVEk0nB8PihR\_0clhxrIbOPE5y\_v6X9PQoqRfFKRbZDiF7o#](https://docs.google.com/document/d/1fbfpR5ZQiVTZYChzWut06fkXA9CQziMMJNtFOiF6Giw/edit?fbclid=IwAR3MU3OTehQ_nVEk0nB8PihR_0clhxrIbOPE5y_v6X9PQoqRfFKRbZDiF7o)

# 2. Materials and methods

## 2.1 Pipeline

How this replication follows pipeline (short overview of methods section)

The replication of this paper follows and provides an exemplification of the steps the pipeline consists of (fig. 1). The rest of the methods section will provide a detailed description of the course of action taken to replicate the paper by \* Chakraborty et al. \*. However, to provide an overview of the process and showcase how it followed our proposed pipeline, a short summary will be provided a long with a figure which visualizes the process (fig. 2).

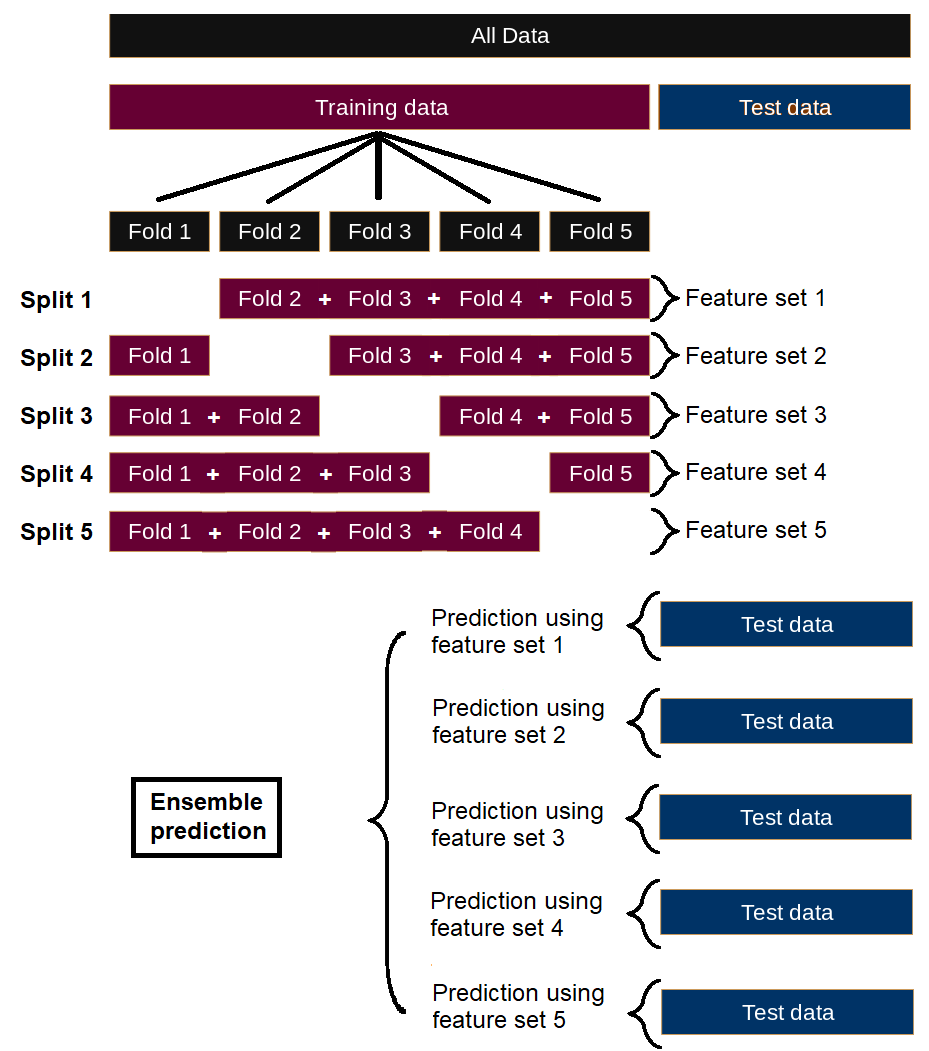


Fig. x \*.

*An overview visualization of this specific machine learning process - partitioning, feature reduction sets and predictions of the holdout set*

First and foremost, cleaned voice data was acquired from which relevant features were extracted (the feature list was in this replication determined by the original study). The voice data was then partitioned into a training and a holdout set in an 80/20 partitioning process that kept IDs separate in each partition. Feature scaling was then performed using min/max normalization such that it did not allow data leakage from the training to the holdout set. As an ensemble model was desired, we then performed feature selection using LASSO on 5 different splits of the training data. The 5 resulting feature lists where then used as parameters for training 5 distinct SVM linear kernel classification models. These model in turn had their parameters tuned before being tested on a test portion of the training data that the models had not seen. After tuning the parameters until a desired outcome was found, the training data was used to train 5 models, each with its own feature set and subsequently predicted the holdout set. The majority vote of a given holdout voice file was then recorded and this constituted the ensemble models predictions. Finally, all models were evaluated based on their performance, with appropriate metrics.

## 2.2 Literature search and choice of replication

A literature search for papers, dissertations and unpublished manuscripts was conducted for finding the paper to replicate. The complete list of papers listed in the meta-analysis by Alberto et al. in 2019 (Alberto et al., 2019) was manually screened – first by title and since by content. As their search was last updated as of April 12 2018, the search was continued from that date and forward in time by the use of search using Google Scholar on the Sep 15 2020, using the same search terms (schizo\* AND machine learning AND prosody OR inflection OR intensity OR pitch OR fundamental frequency OR speech rate OR voice quality OR acoustic OR intonation OR vocal).

The manual search explored the papers by the author, looking for papers that 1) were transparent and well-documented, 2) were thorough in applying proper machine learning methods, 3) had larger amounts of data. The study by Chakraborty et al. from 2018 was chosen for replication on after taking these factors into consideration (Chakraborty et al., 2018).

\* Expand on this? \*

## 2.3 Data

### 2.3.1 Data sources

The data used in this paper consists of speech recordings gathered from 3 published studies (Beck et al., 2020; Bliksted et al., 2014, 2019) and an unpublished study by Vibeke Bliksted.   
Although the data was acquired in separate studies the speech data has several qualities which makes it suitable for combining into a single study:

Participants from all studies went through the same tasks; namely the Frith Happé animations task (FHA) (Abell et al., 2000). All participant went through 8 such trials, except for in the study from 2015 by Bliksted et al., where the participants were presented with 10 trials (Bliksted et al., 2014).

Moreover, recording equipment and recording setting was constant within study, but unique across studies. This results in data corpora of diverse speech recordings suitable for testing whether implementation of a certain machine learning algorithm proves to be versatile in its predictions across data sets.

### 2.3.2 Participants

222 Danish participants were included in this study. Out of the 222 participants 106 were clinically diagnosed with schizophrenia by trained psychiatrists in accordance with the standards of ICD-10 DCR (Zivetz, 1992). Patients were recruited through OPUS, Clinic for people with schizophrenia, Aarhus University Hospital Risskov.  
The patient group was originally matched one-to-one with healthy control subjects (N = 116), using the following criteria: age, sex, handedness, ethnicity, community of residence and parental social economic status (based on the highest parental education and expected parental income according to Statistics Denmark regarding wages) and educational level (based on the last commenced education) (*Statistics Denmark*, n.d.). Healthy control subjects were recruited via advertisements in four local newspapers. All participants in this group (and their first-degree relatives) had no history of any psychological disorders. Although the control group was originally matched one-to-one with the patient group, 14 patients and 4 controls were excluded due to poor recording quality or other similar factors. This explains the uneven number of participants within each group. For further information on participants, see table x \*.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Study | N() | Diagnosis | N(Females) | N(Males) | Mean(Age) | SD(Age) | Range(Age) |
| Beck et al., 2020 | 70 | SZ | 16 | 18 | 22.8 | 3.13 | 18-31 |
| TD | 17 | 19 | 22.7 | 3.19 | 18-30 |
| Bliksted et al., 2014 | 46 | SZ | 6 | 17 | 23.3 | 3.94 | 18-33 |
| TD | 7 | 16 | 23.7 | 3.61 | 18-34 |
| Bliksted et al., 2019 | 48 | SZ | 11 | 8 | 40.8 | 12.4 | 20-61 |
| TD | 13 | 16 | 37.5 | 13.1 | 21-62 |
| Bliksted et al., n.d. | 58 | SZ | 12 | 18 | 24.8 | 3.66 | 18-31 |
| TD | 13 | 15 | 24.4 | 4.65 | 18-34 |
| Total | 106 | SZ | 45 | 61 | 26.7 | 9.02 | 18-61 |
| 116 | TD | 50 | 66 | 26.7 | 9.22 | 18-62 |

Table x \* :

*Demographic data on the sex and diagnosis within each of the different studies. N means number and SD standard deviation.*

### 2.3.2 Procedure/task

The participants went through the Frith Happé animations task. This task consisted of watching a 2D top-view video of animated triangles. There were two distinct triangles; one large red and one small blue, both of which moved around on the screen and most videos furthermore contained an enclosure in the center of the video. There were three conditions with multiple videos for each condition:

**1. Random movement sequences.** There was no obvious interaction between the triangles and movement appears random. **2. Goal-directed (G-D) movement sequences.** An interaction between the triangles in which actions are directed toward each other in order to achieve specific goals.

**3. Mental interaction (ToM)**. An interaction between the triangles involving the manipulation of the emotions and thoughts of one triangle by the other. After watching an animation from one of these conditions, the participants were interviewed and asked to describe what happened in the animation. Each description of a trial thus ended up as a single .wav file.

## 2.4 Preprocessing

### 2.4.1 Cleaning of audio files

The cleaning of the audio files was carried out by Ludvig Olsen in 2018 (Olsen, 2018)  
The audio files were then converted to 16-bit .wav files, with a sample rate of 16k. They were subsequently denoised by stacking multiple instances of the Voice De-noise and De-hum tools in the iZotope RX 6 audio editor (iZotope Inc., 2018). A small equalizer tilt was applied at 1085Hz with the Fabfilter Pro-Q2 equalizer to bring more brightness to the signal (FabFilter Software Instruments, 2018). The signal was normalized to peak at -1dB both before and after the cleaning steps.

### 2.4.2 Feature extraction from audio files

The toolkit openSMILE 2.3.0 was used for extracting the features needed for the SVM classification algorithm. From within the openSMILE software package, the base-set configuration file of emotion recognition features called ‘emobase’ was chosen for feature extraction.

The feature set specified by emobase contains 988 features used for emotion recognition:

Intensity, Loudness, 12 MFCC’s, F0 Pitch, Probability of voicing, F0 envelope, 8 LSFs (Line

Spectral Frequencies), Zero-Crossing Rate. Delta regression coefficients are then computed from all these previously mentioned low-level descriptors (LLD). Both the LLDs and their delta coefficients are smoothed by a moving average window that filters with a window size of 3 seconds. Furthermore, the following functionals are applied to the LLDs and the delta coefficients:

Max./Min. values and their respective relative position within input, range, arithmetic mean, 2 linear

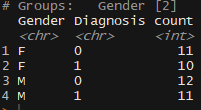
regression coefficients and linear and quadratic error, standard deviation, skewness, kurtosis,

quartile 1-3, and 3 inter-quartile ranges.

This results in the feature set consisting of 988 features. In other words; 26 LLDs, a delta regression coefficient for each LLD and 19 functionals for each of the LLDs and for each of the delta regression coefficients (26 \* 2 \* 19 = 988). The process of feature extraction was executed on each of the speech recordings, yielding a single feature vector for each trial of each participant.

### 2.4.3 Partitioning

To be able to evaluate the performance of the model the dataset was partitioned into a training set and a test set consisting of 80% and 20% of the total data, respectively. The partitioning was carried out using the package groupdata2 and was done semi-randomly (Olsen, 2020). The partitioning kept each participant ID only within one of the two resulting training and test sets. This prevented leakage of information from the training set to the test set, which otherwise would have led to overfitting and finally unprecise values for the evaluation. Moreover, to avoid a skewed distribution of sex or diagnosis between sets (e.g. ending up with only males/controls in the test set as a result of a random partitioning), sex and controls/patients were evenly distributed in the partitioning.



### 2.4.4 Normalization

All feature parameters were normalized using the min-max feature scaling formula in order to achieve a dataset with a common scale without losing information or distorting differences in the range of values.



To avoid overfitting as a result of carrying data from the test set to the training set, the normalization was carried out separately for the training and the testing set. The scaling used the min. and the max. value for each feature, only from the training set, both for the training and for the testing set. This had the advantage of having both the training and the test features on the same scale, while not letting information from the test set flow to the training set and is common practice when applying most machine learning algorithms.

## 2.5 Feature selection using LASSO

### 2.5.1 Motivation for using LASSO

As the 988 acoustic features from the ‘emobase’ package were originally designed to distinguish emotions from speech, some of the features were bound to be redundant for the purpose of distinguishing between patients and controls. As a measure to counterfeit this, a rigorous feature selection method was applied to rid the model of superfluous features. This was done in order to simplify the model and thereby reduces both complexity, computational power needed to run the model and in order to improve both predictive power and interpretability of the classifier.

Feature selection was done using L2 regularization, also called the Least Absolute Shrinkage and Selection Operator (LASSO) analysis regression. To carry out this process, the ‘glmnet’ R Package was utilized for the purpose of this paper. (Friedman et al., 2010)

Although the parameters could have been regularized using Ridge or ElasticNet, LASSO regularization has the advantage of being able to shrink irrelevant parameters all the way to zero – as opposed to Ridge regularization. Elastic net is a combination of Ridge and Lasso and would therefore be a compromise between the two. The shrinking of parameter estimates to zero is beneficial given the many features that are unrelated to the distinction between schizophrenia and healthy individuals.

### 2.5.2 What is L2 regularization?

This method optimizes beta estimates for all parameters not only through misclassification error but also adding a L2 regularization term. The latter adds a penalty to each beta estimate on the basis of a lambda value multiplied with the beta estimate.

In other words; performing L2 regularization means fitting a LASSO regression model and thus finding the optimal beta values for all parameters using the loss function seen below.

The loss function used for finding parameter estimates using LASSO:  

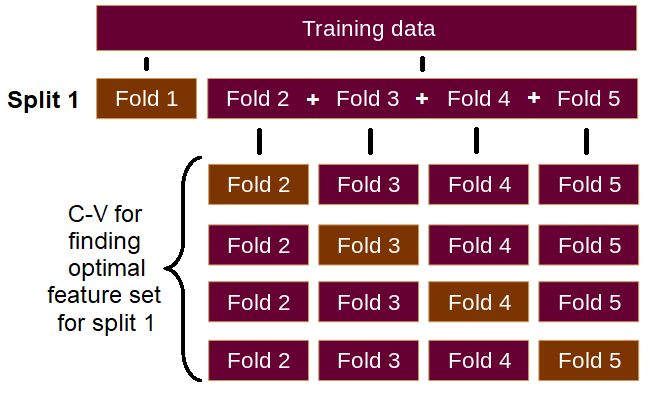

Since this method requires a lambda value (λ), the optimal lambda value also had to be found. The lambda value producing the minimum value in the loss function (lambda.min) was first computed. This was done by testing a range of lambda values using Leave-One-Out CV\*\*\*\* 5-FOLD? SEE FIGURES \*\*\*\*\*\*\* (LOO-CV). Subsequently the lambda value resulting in the fewest number of parameters within 1 SE from the lambda.min was chosen (lambda.1se). Although lambda.min has the lowest level of misclassification, lambda.1se has the advantage of acknowledging the fact that the fits are estimated with some error (Friedman et al., 2010). This process thus generates a list of parameter estimates. Those that have not been shrunken to zero are selected as relevant features for predicting patients from controls. For a visualization, see fig. x \*.



Fig. x \* :

*A range of lambda values (x-axis) and the resulting 1) misclassification error, and 2) number of features (seen at the top). From left to right, the dotted lines represent lambda.min and lambda.1se, respectively.*

### 2.5.4 Feature selection

The training data was partitioned into 5 folds, and thus also 5 splits (see fig. 2). The previously mentioned L2 regularization was carried out on 4/5th’s of each of these splits, resulting in 5 different feature sets (see appendix x\* for list of these feature sets). An illustration of the feature selection for a single split, can be seen below.   
Figure x \* :

*Figure showing the process of feature selection for one of the splits:*

*The training data is divided up into 5 folds. One fold is then excluded. Using cross-validation, the LASSO regression fit for a specific lambda value is then computed with each of the folds being omitted once. The misclassification error for each of these fits is then accumulated and stored. The process is then reiterated using a new lambda value from the lambda grid, until all errors from all relevant lambda values have been obtained. This entire procedure is then repeated for each of the 5 splits.*

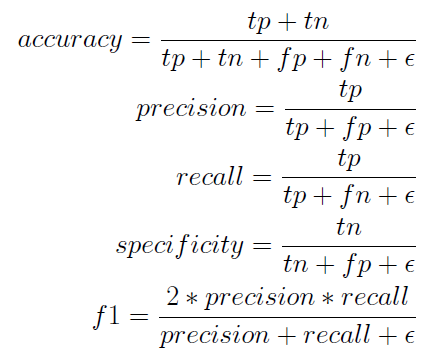
## 2.6 Model and model tuning

**Actual paper:**  
  
**Notes for section:**

1. What do we have now? (short summary of previous sections)
   1. 5 feature lists
   2. Predicting .wav file and not participants
2. Creating the submodels
   1. SVM, linear kernel
3. Testing the submodels
   1. Default search patterns for C and Gamma parameters
   2. Evaluating (through F1-scores as described below)
4. Creating the ensemble model
   1. Decision based on majority vote

## 2.7 Evaluation metrics

**Actual paper:**The predictions on the different models on diagnosis were evaluated on their classification performance. The evaluation metrics used can be seen below:



Where,

TP is True Positives (patients classified as patients),

TN is True Negatives (controls classified as controls),

FP is False Positives (controls classified as patients)

FN is False Negatives (patients classified as controls)

**Notes:**

AUC should be equivalent to F1-score when they’re balanced

- Mention why not much more is needed

Providing confusion matrix

# 3. Results

**Actual paper:**  
**Notes for section:**

This section presents the performance of the machine learning models when predicting various parts of the full data. A crude overview of the performance of the 5 models on the various test sets is given in table 2. An in-depth look at the ensemble models performance; both for controls and for the patient group is provided in table 3. The latter also provides insight into performance differences between the sexes. Finally, confusion matrices (table 4, 5 and 6) provide the necessary details that would underlie calculations for any and all additional performance metrics.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Testing set** | **Training and feature set** | **Macro avg. F1-score** | **Accuracy** | **Baseline accuracy** |
| Train 1 | Train 1 | 0.896 | 89.64% | 53.05 |
| Train 2 | Train 2 | 0.930 | 93.03% | 51.52 |
| Train 3 | Train 3 | 0.897 | 89.73% | 52.21 |
| Train 4 | Train 4 | 0.899 | 89.91% | 51.89 |
| Train 5 | Train 5 | 0.898 | 89.85% | 51.80 |
|  | | | | |
| Test 1 | Train 1 | 0.687 | 68.68% | 51.85 |
| Test 2 | Train 2 | 0.630 | 63.05% | 54.34 |
| Test 3 | Train 3 | 0.678 | 67.84% | 51.62 |
| Test 4 | Train 4 | 0.613 | 61.31% | 52.94 |
| Test 5 | Train 5 | 0.658 | 65.80% | 53.29 |
|  | | | | |
| Holdout | Train 1 | 0.644 | 64.44% | 51.87% |
| Train 2 | 0.652 | 65.19% | 51.87% |
| Train 3 | 0.735 | 73.51% | 51.87% |
| Train 4 | 0.740 | 74.05% | 51.87% |
| Train 5 | 0.716 | 71.64% | 51.87% |
| **Ensemble (majority vote of set 1:5)** | **0.703** | **70.32%** | **51.87%** |

Table 1:

*Prediction performance for all 5 SVM linear kernel models, on various testing data.*

*Within-sample prediction performance can be seen in the first 5 rows, while 5-10 depicts performance tested on the 5 test sets. Finally, the performance for the models’ predictions on the holdout set and the majority decision vote can be seen in the bottommost 6 rows.*

|  |  |  |  |
| --- | --- | --- | --- |
| N = 374 | Predicted group | | |
| True group |  | TD | SZ |
| TD | 130 | 64 |
| SZ | 47 | 133 |

Table 3:  
*Confusion matrix for the ensemble model predictions*

|  |  |  |  |
| --- | --- | --- | --- |
| N = 180 | Predicted group | | |
| True group |  | TD | SZ |
| TD | 64 | 28 |
| SZ | 26 | 62 |

Table 4:

*Confusion matrix for the female ensemble model predictions*

|  |  |  |  |
| --- | --- | --- | --- |
| N = 194 | Predicted group | | |
| True group |  | TD | SZ |
| TD | 66 | 36 |
| SZ | 21 | 71 |

Table 5:

*Confusion matrix for the male ensemble model predictions*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Test set | Model | Sex | Acc. | Baseline acc. | Class | Precision | Recall | F1-score |
| Holdout | Ensemble | Male | 70.62% | 52.58% | SZ | 0.664 | 0.772 | 0.714 |
| TD | 0.759 | 0.647 | 0.698 |
| Female | 70.00% | 51.11% | SZ | 0.689 | 0.705 | 0.697 |
| TD | 0.711 | 0.696 | 0.703 |
| Both | 70.32% | 51.87% | SZ | 0.675 | 0.739 | 0.706 |
| TD | 0.734 | 0.670 | 0.700 |

Table 2:

*Performance of the ensemble model - within both the sexes and diagnosis (TD stands for typically developed/the control group while SZ represents the schizophrenia group).*

# 4. Discussion

## 4.1 Results and replication comparison

### 4.1.1 Performance

**Actual:**

**Notes:**

1. Performance of models on test
   1. F1-score (and short mention of accuracy)
      1. F1-score for model overall
      2. F1-scores for patients and controls respectively
   2. Precision + recall
   3. Between sexes
      1. This: Well-balanced in replication
      2. Original: No information in original paper
         1. Ought to be included
2. Where do the differences in performance come from?
   1. Methods (as will be discussed in next section)

Maybe include:

1. Performance of models on train
   1. High performance
   2. Low generalizability due to overfitting
      1. Mention bad study that overfits

* These things in terms of what is being predicted (train/test/holdout) - better predictions when overfitting – obviously. To underline how much more shit the predictions will get when you’re rigorous. With increased conservatism, how good are the results really?
* Performance between groups
  + Sexes:
    - This: Balanced test-set (but perhaps a bit worse training)
    - Original: No information. But balanced dataset.
  + Nationalities
    - This: None
    - Original: No information. But unbalanced dataset
* Where do the differences come from?
  + Data, Feature selection, Methods
  + See next sections

### 4.1.2 Data

**Actual:**

**Notes:**

* Language/nationality
  + Biased because of difference in labeling
    - This: Danish diagnostics
    - Original: Chinese, Malay, Indian diagnostics
  + Biased because of difference in language
    - This: Danish
    - Original: 3 Countries, with different languages
* Task
  + This: mid-level difficulty; description of triangles. No social component
  + Original: high-level difficulty; interview. Social component
* Data quantity
  + This: More participants with shorter recordings
  + Original: Fewer participants with longer recordings
* Sound quality
  + This: Difference in recording equipment
  + Original: Maybe?
* What contributed to the differences in performance? (If any)
  + Possibly all. Likely not sound quality to a large extent

### 4.1.3 Feature selection

**Actual:**

**Notes:**

1. Type of feature selection
   1. This: LASSO
   2. Original: PCA
2. Hard to replicate, given the sparse information on how PCA was used
   1. Their feature selection method hard to follow
   2. Could have been understood in two different ways
3. Specific feature selection method shouldn’t have a large impact on performance

Explanation of how it could be understood:

“*the features of the training set were ranked using one of the following techniques: F-score (ANOVA), χ 2 , Mutual Information, Pearson correlation, Principal Components, linear SVM, Decision Trees, and Random Forests. Subsequently, the optimal number of features were selected according to the previous ranking methods*”  
PCA used to rank? Most common method is that PCA is used for defining new features, namely PC1 + PC2 + ... +PCn, until some desired threshold of accumulated variance is met.

There’s also the possibility that it truly was used to rank, e.g. by looking at the features with least shared variance in the different principal components to avoid covarying features, but also here it is not possible to replicate 1-1. The method is still not specified

Shouldn’t really matter:

Regardless, of method used by Chakraborty et al, the method used here is good. And if the method using speech for classification truly is robust, then either would work. If these results truly are reliable and reliable, they shouldn’t be dependent on PCA/LASSO / whatever

Understanding PCA notes:

Link of idea of PCA for feature selection. (starts at 3:50). It shows that there are different methods (example with gain, here)

<https://www.youtube.com/watch?v=YEDOSOd44bU&list=PLBv09BD7ez_5_yapAg86Od6JeeypkS4YM&index=2&frags=wn&ab_channel=VictorLavrenko>

Link for example of PCA for feature selection (creating new features):

<https://www.quora.com/How-do-you-use-PCA-for-feature-selection>

### 4.1.4 Methods ()

**Actual:**

**Notes:**

1. Predicting (single participants, or same participants multiple times)
   1. This: Predicting .wav files (several for each participant)
   2. Original: Predicting participants
   3. Does this matter?
2. Ensemble modeling vs. Single machine learning algorithm
   1. Stacking ensemble modeling
      1. Better (if models are diverse, and generally good)
      2. Only very slightly better
   2. Single machine learning algorithm
      1. Slightly worse

MAYBE INCLUDE:

* Specifics on ensemble modeling
  + Diversity/data trade-off in ensemble modeling

FOR BELOW I DO THE OPPOSITE NOW!!!!

Diversity/data trade-off in ensemble modeling:

I use all training data in each of the ensemble-sub-models. As opposed to excluding the test sets, that were also excluded for feature selection.

Could this be an issue?  
Yes; groups of diverse problem solvers (in general) outperform, the best (also often similar) models. At least when the diverse problem solvers and the better, more similar models have roughly the same amount of data. (Hong & Page, 2004)

Why did I choose to do it anyways?  
The “diversity” the opposing idea would bring, is not due to difference in neither type of models or any other diversity parameter. The opposing choice would only give diversity from differences in training data.

The increase in diversity would in this study, be on the cost of less training data. And less training data means worse predictions in general.

Does it really matter?

The trade-off between more/less training data and more/less diversity is unlikely to have had much of an impact. E.g. Less than 2 percent increase in acc. when having 10 agents (and we only have 5, which would probably mean even less of an impact) (Hong & Page, 2004).

But the potentially very small positive effect a more diverse set of decision-agents, might very well be negated by the fact that all of the 5 diverse models would be worse, due to their more limited data. In other words; no – it isn’t likely to have had a large effect. But it would have been interesting to do both.

## 4.2 Pipeline

### 4.2.1 (Narrow) How did an implementation of pipeline in this replication work out?

**Actual:**

**Notes:**

1. Replication
   1. Possible
   2. Hard (Methods explained in condensed manner)
2. Comparison
   1. Possible
   2. Hard (More information on sexes and nationalities needed)
3. Getting similar results
   1. Differences in performance – where does it come from?
      1. Biased labels
      2. Difference in language
      3. Task differences
      4. Difference in algorithms
      5. Arbitrary choices for tuning
      6. A mixture (which mixture?) of all the above
   2. Some things might balance each other’s out, some might not

### 4.2.2 Problems established from conservative replication

**Actual:**

**Notes:**

1. Curious that other studies have found much(!) higher accuracies
   1. Study 1
   2. Study 2
   3. Overfitting?
      1. My predictions on training 90% accuracy
      2. Scaling
2. Hard to know where differences in performance come from
   1. (All the differences on task, data, language, labeling etc.)
   2. Solution: More documentation on this and more reproductions
3. Bad documentation is insufficient for facilitating replication
   1. From practical experience
4. It is up to individual researchers and their experience to produce original studies and replications alike (not good)
   1. Arbitrary choices and handycrafts
      1. Tuning (C-parameters)
      2. Model type
   2. From practical experience – not possible to find established pipeline and solutions

To underline how much more shit the predictions will get when you’re rigorous  
With increased conservatism, how good are the results really?

Results real good if overfit

To underline how much more shit the predictions will get when you’re rigorous  
With increased conservatism, how good are the results really?

**Discussion:**

Many choices have to do with handycrafts and arbitrary choices (tuning)

We haven’t gotten enough research on principles of how to do this (Right now up to individual experience of researchers)

Present the issue that this hasn’t been fixed by our paper either. Even with relatively simple algorithms. Deep learning would mean that this is even worse.

So is language/way of speaking

* Hard to know whether differences in performance are due to:
  + biased labels
  + difference in language
  + task differences
  + difference in machine learning algorithm
  + arbitrary choices for tuning
  + a mixture of all (could balance each other out, if in opposite directions (some make it harder for the Danish corpus, some make it easier)

## 4.3 Further research

### 4.3.1 Need for a widely applicable, conservative, transparent pipeline.

1. Meta (widely applicable, conservative, transparent pipeline) would help by:
   1. Avoiding in overfitting (as mentioned previously)
   2. Making it easier to compare results (as mentioned previously)
      1. Within or across sexes and nationalities,
   3. Making it easier to replicate
   4. Enabling research to know locate the origin of differences in results (as mentioned previously)
      1. Biased labels
      2. Difference in language
      3. Task differences
      4. Difference in algorithms
      5. Arbitrary choices for tuning
      6. A mixture (which mixture?) of all the above
   5. Shedding light on arbitrary choices – either by:
      1. Providing information on it in the papers
      2. Providing a method for making these choices
2. In general: More replications and a generally more open-science based approach

This study was not enough.

**Discussion:**

Many choices have to do with handycrafts and arbitrary choices (tuning)

We haven’t gotten enough research on principles of how to do this (Right now up to individual experience of researchers)

Present the issue that this hasn’t been fixed by our paper either. Even with relatively simple algorithms. Deep learning would mean that this is even worse.

1. Establish a widely accepted pipeline. It should be:
   1. Rigorous
   2. Conservative
   3. Transparent
2. It allows for
   1. Replications
   2. Transparency

To test generalizability and robustness

This allows for more replications

# 5. Conclusion

**Actual paper:**

**Notes:**

# 6. Acknowledgements

s

# 7. References

Abell, F., Happé, F., & Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cognitive Development*, *15*(1), 1–16. https://doi.org/10.1016/S0885-2014(00)00014-9

Alberto, P., Arndis, S., Vibeke, B., & Riccardo, F. (2019). *Voice Patterns in Schizophrenia: A systematic Review and Bayesian Meta-Analysis* [Preprint]. Bioinformatics. https://doi.org/10.1101/583815

Beck, K. I., Simonsen, A., Wang, H., Yang, L., Zhou, Y., & Bliksted, V. (2020). Cross-cultural comparison of theory of mind deficits in patients with schizophrenia from China and Denmark: Different aspects of ToM show different results. *Nordic Journal of Psychiatry*, 1–8.

Bliksted, V., Fagerlund, B., Weed, E., Frith, C., & Videbech, P. (2014). Social cognition and neurocognitive deficits in first-episode schizophrenia. *Schizophrenia Research*, *153*(1), 9–17. https://doi.org/10.1016/j.schres.2014.01.010

Bliksted, V., Frith, C., Videbech, P., Fagerlund, B., Emborg, C., Simonsen, A., Roepstorff, A., & Campbell-Meiklejohn, D. (2019). Hyper-and hypomentalizing in patients with first-episode schizophrenia: FMRI and behavioral studies. *Schizophrenia Bulletin*, *45*(2), 377–385.

Chakraborty, D., Yang, Z., Tahir, Y., Maszczyk, T., Dauwels, J., Thalmann, N., Zheng, J., Maniam, Y., Amirah, N., & Tan, B. L. (2018). Prediction of negative symptoms of schizophrenia from emotion related low-level speech signals. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6024–6028.

FabFilter Software Instruments. (2018). *FabFilter* (Fabfilter pro-q 2.) [Computer software].

Friedman, J., Hastie, T., & Tibshirani, R. (2010). *Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software*. *33(1)*, 1–22.

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, *101*(46), 16385–16389. https://doi.org/10.1073/pnas.0403723101

iZotope Inc. (2018). *IZotope RX 6*.

Olsen, L. (2018). *Automatically diagnosing mental disorders from voice: A deep learning approach*.

Olsen, L. (2020). *groupdata2: Creating Groups from Data* (1.3.0) [Computer software]. https://CRAN.R-project.org/package=groupdata2

*Statistics Denmark*. (n.d.). Retrieved 11 November 2020, from https://www.dst.dk/en/

Zivetz, L. (1992). *The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines* (Vol. 1). World Health Organization.

# 8. Appendix

s